(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 274 012 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
08.01.2003 Bulletin 2003/02

(51) Int Cl.[7]: **G06F 11/00**

(21) Application number: **02254265.8**

(22) Date of filing: **19.06.2002**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **05.07.2001 US 900298**

(71) Applicant: **Sun Microsystems, Inc.**
Santa Clara, California 95054 (US)

(72) Inventor: **Shirriff, Kenneth W.**
Redwood City, CA 94061 (US)

(74) Representative: **Davies, Simon Robert**
D Young & Co,
21 New Fetter Lane
London, EC4A 1DA (GB)

(54) Method and system for establishing a quorum for a geographically distributed cluster of computers

(57) One embodiment of the present invention provides a system that facilitates establishing a quorum for a cluster of computers that are geographically distributed. The system operates by detecting a change in membership of the cluster. Upon detecting the change, the system forms a potential new cluster by attempting to communicate with all other computers within the cluster. The system accumulates votes for each computer successfully contacted. The system also attempts to gain control of a quorum server located at a site separate from all computers within the cluster. If successful at gaining control, the system accumulates the quorum server's votes as well. If the total of accumulated votes is a majority of the available votes, the system forms a new cluster from the potential new cluster.

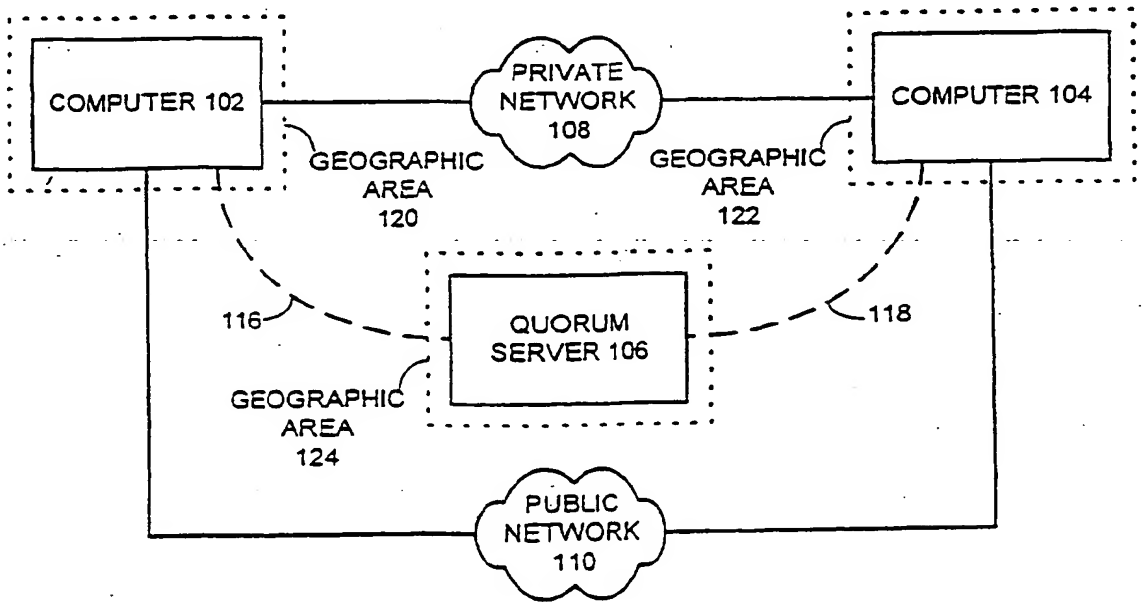

FIG. 1

EP 1 274 012 A2

ter with nodes that are widely separated, by potentially thousands of miles, in order to provide reliability in the event of a local disaster. This separation poses problems for the quorum configuration. If the quorum device is located with either node, a disaster at that site could destroy both the node and the quorum device, effectively preventing the other node from taking control. In addition, connecting a quorum device such as a SCSI disk over these long distances can be extremely expensive or impossible.

[0014] Accordingly, one embodiment of the present invention provides a system that facilitates establishing a quorum for a cluster of computers that are geographically distributed. The system operates by detecting a change in membership of the cluster. Upon detecting the change, the system forms a potential new cluster by attempting to communicate with all other computers within the cluster. The system accumulates votes for each computer successfully contacted. The system also attempts to gain control of a quorum server located at a site separate from all computers within the cluster. If successful at gaining control, the system accumulates the quorum server's vote or votes as well. If the total of accumulated votes comprises a majority of the available votes, the system forms a new cluster from the potential new cluster.

[0015] In one embodiment of the present invention, the system exchanges heartbeat messages with all other computers that are part of the cluster. Upon discovering an absence of heartbeat messages from any computer in the cluster, the system initiates a cluster membership protocol.

[0016] In one embodiment of the present invention, detecting the change in cluster membership includes detecting that the cluster has not been formed.

[0017] In one embodiment of the present invention, attempting to gain control of the quorum server involves communicating with the quorum server using cryptographic techniques.

[0018] In one embodiment of the present invention, the system exchanges a status message with each member of the new cluster. The system updates the local status of the computer to the most recent status available within the status messages.

[0019] Another embodiment of the present invention provides a system that facilitates establishing a quorum for a cluster of computers that are geographically distributed. The system provides a quorum server at a site separate from a location of any computer within the cluster. The system assigns at least one vote to each computer within the cluster. The system also assigns at least one vote to the quorum server. In operation, the system attempts to establish communications between each pair of computers within the cluster. A count of votes is accumulated at each computer for each computer that responds. The system also attempts to establish control over the quorum server from each computer within the cluster. If control is established over the quorum server, the quorum server's vote(s) are accumulated in the count of votes. The system establishes a quorum when a majority of available votes has been accumulated in the count of votes.

[0020] In one embodiment of the present invention, the quorum server grants control to only a first computer attempting to establish control. Another approach is for the quorum server to grant control to only one computer out of all the computers attempting to establish control based on a pre-established priority list.

[0021] In one embodiment of the present invention, votes are assigned so that the quorum includes at least one computer that was in an immediately previous cluster. This ensures that a cluster formed from the quorum has current data.

[0022] In one embodiment of the present invention, attempting to establish control over the quorum server involves establishing communications with the quorum server. Note that cryptographic techniques may be employed here to deter attacks.

[0023] Various embodiments in accordance with the invention will now be described in detail by way of example only, with reference to the following drawings:

FIG. 1 illustrates a geographically distributed cluster of computers coupled together in accordance with one embodiment of the present invention.

FIG. 2 is a flowchart illustrating the process of detecting and processing a failure within a cluster in accordance with one embodiment of the present invention.

FIG. 3 is a flowchart illustrating the process of determining cluster membership, such as may be used in the process of FIG. 2.

FIG. 4 is a flowchart illustrating the process of granting control of a quorum server such as shown in FIG. 1.

FIG. 5 is a flowchart illustrating the process of reconfiguring a computer within a cluster, such as may be used in the process of FIG. 2.

Computer Cluster

[0024] FIG. 1 illustrates a geographically distributed cluster of computers coupled together in accordance with one embodiment of the present invention. Computers 102 and 104 form a cluster of computers that operate in concert to provide services and data to users. Two or more computers are formed into a cluster to provide speed and reliability for the users. Computers 102 and 104 are located in geographic areas 120 and 122 respectively. Geographic areas 120 and 122 are widely separated, possibly by thousands of miles, in order to provide survivability for the cluster in case of a local disaster at geographic area 120 or 122. For example, geographic area 120 may be located in California, while geographic area 122 may be located in New York.

[0025] Computers 102 and 104 can generally include

nodes of the membership of the new cluster (step 316). Note that the above steps are being accomplished by all computers in the system simultaneously.

Controlling Quorum Server

[0039] FIG. 4 is a flowchart illustrating the process of granting control of quorum server 106 in accordance with one embodiment of the present invention. The system starts when quorum server 106 receives a request for control from a node in the proposed new cluster (step 402). Next, quorum server 106 determines if the requesting node was on the list of nodes for the previous cluster (step 404). If the requesting node was not on the list of nodes for the previous cluster, quorum server 106 determines if the list of nodes for the previous cluster is empty (step 406). Note that an empty list indicates that a cluster had never been formed and this request is part of initializing a cluster for the first time. If the cluster list is not empty at 406, quorum server 106 denies the request to control quorum server 106 (step 408).

[0040] If the node was on the previous cluster list at 404 or if the cluster list is empty at 406, quorum server 106 sets the cluster list to contain only the requesting node (step 410). (It will be apparent to a person of ordinary skill in the art that there are other ways to reset the list, including receiving a list of nodes from the requesting node to include in the list, or receiving a list of nodes from the requesting node to exclude from the list). Finally, quorum server 106 affirms the request to control quorum server 106 and grants its vote(s) to the requesting node (step 412).

Reconfiguring a Computer

[0041] FIG. 5 is a flowchart illustrating the process of reconfiguring a computer within a cluster in accordance with one embodiment of the present invention. The system starts when a computer, say computer 102, receives status data from other nodes in the new cluster (step 502). Next, computer 102 determines which set of status data is the most recent (step 504).

[0042] Computer 102 updates its own internal status to conform with the most recent status data available (step 506). Finally, computer 102 informs quorum server 106 which nodes to include in the new cluster list (step 508).

[0043] The data structures and code described herein for implementing the establishment of a quorum are typically stored on a computer readable storage medium, which may be any device or medium that can store code and/or data for use by a computer system. This includes, but is not limited to, magnetic and optical storage devices such as disk drives, magnetic tape, CDs (compact discs) and DVDs (digital versatile discs or digital video discs), and computer instruction signals embodied in a transmission medium (with or without a carrier wave upon which the signals are modulated). For example, the transmission medium may include a communications network, such as the Internet.

[0044] The foregoing description of various embodiments of the present invention has been provided in the context of a particular application, and for the purpose of illustration only. Many other modifications and variations will be apparent to practitioners skilled in the art, and so the scope of the present invention is not limited to the particular embodiments shown, but rather is defined by the appended claims and equivalents thereof.

Claims

1. A method for facilitating the establishment of a quorum for a cluster within a plurality of computers that are geographically distributed, the method comprising the steps of:

   detecting a change in membership of the cluster at a computer within the plurality of computers; and
   upon detecting the change in membership, forming a potential new cluster by attempting to communicate with
   all other computers within the plurality of computers, accumulating votes for each computer successfully contacted,
   attempting to gain control of a quorum server located at a site separate from all computers within the plurality of computers,
   if successful, accumulating the quorum server's votes, and
   if the total of accumulated votes represents a majority of the available votes, forming a new cluster from the potential new cluster.

2. The method of claim 1, wherein the step of detecting a change in membership includes the steps of:

   exchanging heartbeat messages with all computers that are part of the cluster; and
   upon discovering an absence of a heartbeat message from any computer in the cluster, initiating a cluster membership protocol.

3. The method of claim 1, wherein the step of detecting the change in cluster membership includes detecting that the cluster has not been formed.

4. The method of any preceding claim, wherein the step of attempting to gain control of the quorum server includes communicating with the quorum server using cryptographic techniques.

5. The method of any preceding claim, further comprising the steps of:

15. A system to facilitate establishing a quorum for a cluster within a plurality of computers that are geographically distributed, wherein the plurality of computers are coupled together by a network, the system comprising:

    a quorum server located at a site separate from any one computer of the plurality of computers; and

    an independent communications link for coupling each computer of the plurality of computers to the quorum server.

16. The system of claim 15, wherein the quorum server includes a mechanism for granting control to only one computer of the plurality of computers requesting control.

17. The system of claim 15, wherein the quorum server includes a mechanism for maintaining a list of computers accepted into the cluster.

18. The system of any of claims 15 to 17, wherein the quorum server includes a mechanism for cryptographically ensuring an identity of a computer attempting to establish control.

19. The system of any of claims 15 to 18, wherein the quorum server includes monitoring means to monitor the status of each computer within the plurality of computers.

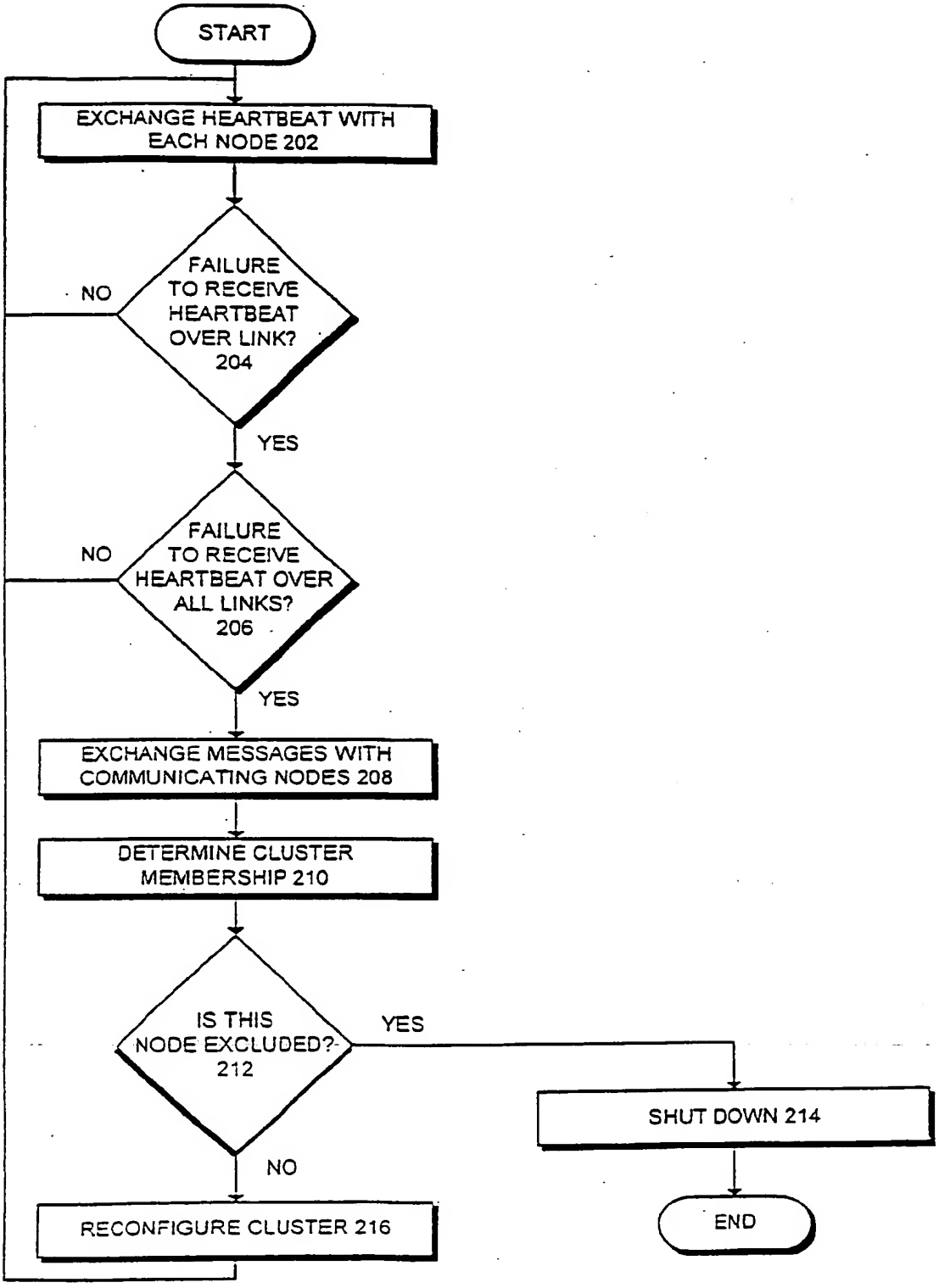

FIG. 2

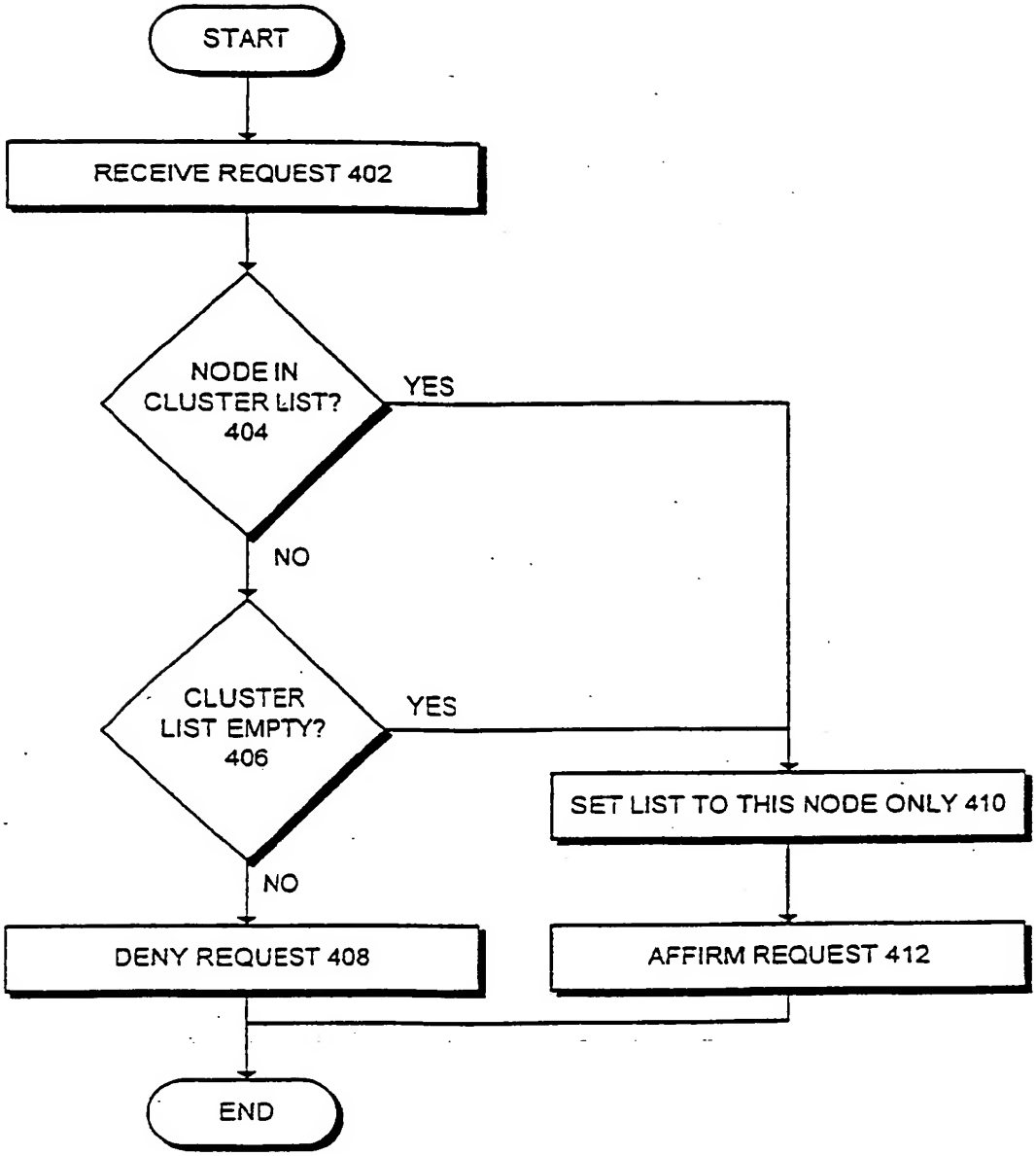

FIG. 4

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 274 012 A3

(12) EUROPEAN PATENT APPLICATION

(88) Date of publication A3:
27.07.2005 Bulletin 2005/30

(43) Date of publication A2:
08.01.2003 Bulletin 2003/02

(21) Application number: **02254265.8**

(22) Date of filing: **19.06.2002**

(51) Int Cl.[7]: **G06F 11/00**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **05.07.2001 US 900298**

(71) Applicant: **Sun Microsystems, Inc.**
Santa Clara, California 95054 (US)

(72) Inventor: **Shirriff, Kenneth W.**
Redwood City, CA 94061 (US)

(74) Representative: **Davies, Simon Robert**
D Young & Co
120 Holborn
London, EC1N 2DY (GB)

(54) Method and system for establishing a quorum for a geographically distributed cluster of computers

(57) One embodiment of the present invention provides a system that facilitates establishing a quorum for a cluster of computers that are geographically distributed. The system operates by detecting a change in membership of the cluster. Upon detecting the change, the system forms a potential new cluster by attempting to communicate with all other computers within the cluster. The system accumulates votes for each computer successfully contacted. The system also attempts to gain control of a quorum server located at a site separate from all computers within the cluster. If successful at gaining control, the system accumulates the quorum server's votes as well. If the total of accumulated votes is a majority of the available votes, the system forms a new cluster from the potential new cluster.

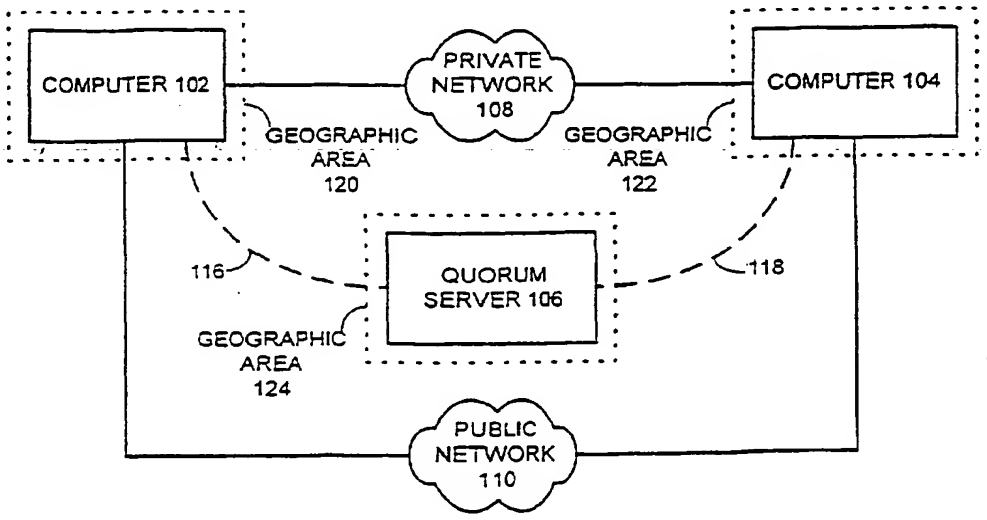

FIG. 1

EP 1 274 012 A3

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 02 25 4265

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
A	RAJAGOPALAN B: "Membership protocols for distributed conference control" COMPUTER COMMUNICATIONS, ELSEVIER SCIENCE PUBLISHERS BV, AMSTERDAM, NL, vol. 18, no. 10, October 1995 (1995-10), pages 695-708, XP004032441 ISSN: 0140-3664 * page 698, column 1, paragraph 4 - page 703, column 1, paragraph 3 *	1-19	
A	PATENT ABSTRACTS OF JAPAN vol. 2000, no. 21, 3 August 2001 (2001-08-03) & JP 2001 117895 A (INTERNATL BUSINESS MACH CORP <IBM>), 27 April 2001 (2001-04-27) * abstract * & US 6 542 929 B1 (BRISKEY KENNETH C ET AL) 1 April 2003 (2003-04-01) * column 2, line 11 - line 56 *	1-19	
A	US 5 999 712 A (MOIIN ET AL) 7 December 1999 (1999-12-07) * column 5, line 9 - column 6, line 29 *	1-19	TECHNICAL FIELDS SEARCHED (Int.Cl.7)

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
Munich	2 June 2005	Bozas, I

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)